Resource-Constrained Sampling of Multiple Compressed Videos

Related Application

[01] This is a continuation-in-part application of U.S. Patent Application Sn. 10/639,951, "Resource-Constrained Encoding of Multiple Videos," filed by Vetro et al. on August 13, 2003.

Field of the Invention

[02] The invention relates generally to sampling compressed videos, and more particularly to the sampling compressed videos according to resource constraints.

Background of the Invention

[03] Encoding of multiple videos has been considered in two major areas: transmission and recording. Transmission of multiple videos is mainly applied in broadcasting and delivery applications, while recording of multiple videos is usually applied to surveillance applications. Although video recording is common in many consumer electronics products, such video recorders deal typically with the encoding of a single video, rather than multiple concurrent videos.

[04] In television broadcasting applications, it is common practice to encode multiple videos, such that the encoded video bitstreams can be transmitted together over a single channel having a fixed bandwidth. For instance, given *N* programs and a total channel bandwidth of 45 Mbps, which is common for satellite links, the problem is to encode the *N* programs with an overall maximum quality, and

multiplex them onto the single channel. Because the bandwidth is fixed and the complexity of each program varies, each of the programs is encoded at a variable bit-rate (VBR). In this way, a near-constant distortion can be maintained across all programs. Thus, more complex portions of one videos can be allocated more bits, by decreasing the bits allocated for less complex portions of other videos that are concurrently encoded.

[05] The encoding process described above is referred to as statistical multiplexing. Techniques associated with this process are described by Haskell in "Multiplexing of Variable Rate Encoded Streams," IEEE Transactions on Circuits and Systems for Video Technology, 1994, Wang et al, "Joint Rate Control For Multi-Program Video Coding," IEEE Transactions on Consumer Electronics, Vol. 42, No. 3, August 1996, U.S. Patent No. 6,091,455, "Statistical Multiplexer for Recording Video," issued to Yang on July 18, 2000, U.S. Patent No. 6,195,388, "Apparatus and Method for Encoding Multiple Video Programs," issued to Choi et al. on February 27, 2001, and references included therein.

[06] Along similar lines, Sun and Vetro have described the encoding of multiple objects in a scene subject to a fixed bandwidth constraint in U.S. Patent No. 5,969,764, issued on October 19, 1999. That method allocates bits to each object. In U.S. Patent Application Sn. 09/579,889, "Method for encoding and transcoding multiple video objects with variable temporal resolution," filed by Vetro et al. on May 26, 2000, a method to satisfy a total bandwidth constraint with each object in a scene having a different temporal rate is described. There, the method minimizes composition artifacts that occur when multiple objects in a scene are encoded at different temporal rates.

[07] The above prior art methods encode multiple videos or objects subject to a *total* bandwidth constraint of a *single* transmission channel.

[08] In the prior art, resource constraints other than bandwidth have been considered in the processing of multiple videos, see for example, U.S. Patent No. 6,052,384, "Using a Receiver Model to Multiplex Variable Bit-Rate Streams Having Timing Constraints," issued to Huang et al. on April 18, 2000, which describes techniques to determine the output bit rates of each stream so that neither the queue for the bitstream in the multiplexer nor the buffer in a decoder overflows or underflows. The rates are determined using timing information that is read from the bitstream, and a receiver model that considers the operation of the decoder buffer.

[09] Transcoding multiple videos considering timing, both delay and processing, constraints are described in U.S. Patent No. 6,275,536, "Implementation architectures of a multi-channel MPEG video transcoder using multiple programmable processors," issued to Chen et al. on August 14, 2001. Input bitstreams are first partitioned into processing units. In one architecture, the processing units are split into different substreams; each substream with its own queue, then each substream is processed in a corresponding branch. In a second architecture, the processing units are assigned to any available processor from a common queue. Independent processing units are processed concurrently according to a queuing system model to minimize an averaging processing time. In contrast to the first architecture that is a parallel process of multiple branches, the second architecture is a single branch to multi-processing.

[010] Similar to Chen et al., U.S. Patent No. 6,008,848, "Video Compression Using Multiple Computing Agents," issued to Tiwari et al., on December 28, 2001 also describes a system and method that uses multiple processors. In contrast, Tiwari applies to encoding of a video and describes techniques to achieve the encoding using coarse grain parallelism effected by multiple processors or compressing agents.

[011] Figure 1 shows a general system model for encoding multiple videos for a surveillance application. Cameras 101 acquire videos 102 for a video recorder 110. Typically, the recorder 110 compresses the videos. The compressed videos are then stored in a memory 120. Later, a video player 130 can play the stored videos.

[012] Figure 2 shows the details of the recorder 200. The acquired videos 102 are sent to a high-speed switch 210. The switch samples the analog video signals in time. The samples are fed to a decoder 220, and the digitized images are encoded by a still-image-encoder 230 to yield compressed images. A memory controller 240 writes the compressed images to allocated space in the memory 120. The stored video can be played back later.

[013] The main problem with the recorder of Figure 2 is that still images are encoded. That does not exploit any temporal redundancy in the videos. As a result, the memory 120 needs to be very large if weeks and months of surveillance videos are stored. It should be noted that a huge amount of surveillance videos, particularly those taken at night, are of totally static scenes. Significant events are rare.

[014] Figure 3 shows an obvious solution to the above problem. In this scheme, there is one encoding channel for each video. In the encoding channel, the video is first NTSC decoded 220. Due to the predictive encoding used in video coders, such as MPEG, a frame memory 310 is maintained for each video encoder 320 to store reference pictures. The input frames from the different camera 101 are then encoded separately and the results are written to the memory 120 using the memory controller 240. The temporal rate of the input frames can be controlled by uniformly sampling with fixed period *T* 301. This sampling makes better utilization of the memory 120. The main drawbacks of that scheme are that the video encoders 320 are not fully utilized considering that they are typically designed to handle full-rate video. Also, the many decoders and encoders increase the cost of the system.

[015] In U.S. Patent No. 6,314,137, "Video data compression system, video recording/playback system, and video data compression encoding method," issued to Ono et al. on November 6, 2001, a system and method that overcomes the above drawbacks is described, as shown in Figure 4. There, a single video encoder 420 is used to encode all of the videos. The digitized video frames from each camera input 101 are sub-sampled with period *T* and buffered in the respective frame memories 210. In order to achieve the predictive encoding with the single encoder 420, a series of input video frames corresponding to one camera input are fed into the video encoder so that predictive coding from frames of the same camera input can be made successively. The Group of Pictures (GOP) structure in MPEG coding allows independent units of such to be formed. In that way, the memory controller 240 becomes a GOP-select, and the GOP's from each camera input are time-multiplexed into the encoder according to the controller 410. With that scheme, a *single* bitstream for all camera inputs is produced. To identify the

portions of the videos that correspond to a given camera, a camera identifier 401 is multiplexed into the encoded bitstream.

[016] With the above solution, one GOP's worth of data is required to be stored in each of the frame memories, which is a much larger requirement than the system of Figure 2 that only requires 1 or 2 reference pictures, at most. Therefore, although there is a significant savings in encoding hardware, memory requirements are still large and expensive. This drawback cannot be overcome by simply sampling the input video more aggressively. Although this will reduce the temporal resolution of the video, the same data for a GOP still needs to be buffered. Only a shorter GOP period would reduce the memory requirements, but this would imply more frequent intra frame coding, which means less coding efficiency. In the most extreme case, a GOP period of 1 would degenerate to the still image coding system shown in Figure 2.

[017] High memory requirements are just one drawback of the system in Figure 4. The problem becomes proportionately worse when the system is scaled to higher number of videos.

[018] In large-scale systems, compressed videos are often used to reduce bandwidth and storage. Therefore, it is desired to provide a system and method for concurrently sub-sampling multiple compressed videos.

Summary of the Invention

[019] One object of the invention is to provide a low-cost, compression-efficient, scalable and flexible system for storing multiple videos with constrained resources, particularly memory, encoding hardware, and delay.

[020] A further object of the invention is to provide a method for achieving variable and non-uniform temporal resolution among multiple uncorrelated compressed videos.

[021] A method acquires compressed input videos. The compressed frames of each input video are acquired at a fixed sampling rate.

[022] Joint analysis is applied concurrently and in parallel to the compressed videos to determine a variable and non-uniform temporal sampling rate for each compressed video so that a combined distortion is minimized and a combined frame rate constraint is satisfied.

[023] Each compressed video is then sampled at the associated variable and non-uniform temporal sampling rate to produce output compressed videos having variable temporal resolutions.

Brief Description of the Drawings

[024] Figure 1 is a block diagram of a prior art recording and playback system for multiple videos;

[025] Figure 2 is a block diagram of a prior art still image encoding system for multiple videos;

[026] Figure 3 is a block diagram of a prior art video encoding system for multiple videos that utilizes multiple encoders;

[027] Figure 4 is a block diagram of a prior art video encoding system for multiple videos that utilizes a single encoder;

[028] Figure 5 is a block diagram of concurrent full frame rate input videos analyzed to produce output videos with non-uniform temporal resolution according to the invention;

[029] Figure 6A is a video encoding system for multiple videos that utilizes a single encoder and individual memories in accordance with the invention;

[030] Figure 6B is a video encoding system for multiple videos that utilizes a single encoder and shared memory in accordance with the invention;

[031] Figure 7 is a block diagram of a video encoding system for multiple videos that utilizes a single encoder and transcoder in accordance with the invention;

[032] Figure 8 is a block diagram of a video processing system operating on intra-coded frames of compressed videos; and

[033] Figure 9 is a block diagram of a video processing system operating on inter-coded frames of compressed videos.

Detailed Description of the Preferred Embodiment

[034] Figure 5 shows individual frames 501 of multiple input videos (videos 1-4) 510 acquired concurrently with a camera at fixed sampling rates. Subject to joint analysis 600 according to our invention, frames 502 of multiple output videos 520 have a variable and non-uniform temporal sampling rates. The objective of the joint analysis is to minimize a combined distortion for all output videos, while a combined frame rate constraint is satisfied.

[035] Factors affecting the rate and presence of a particular output frame 502 at a given time include compression efficiency, resource constraints, and significant event detection. Compression efficiency can relate to the type encoder that is used, e.g., MPEG-2 or MPEG-4. Resource constraints consider memory, processing speed, and encoding rates. By significant events, we mean short-term local events that are different from longer portion of the underlying scene.

[036] For example, in a surveillance video, a person entering an otherwise static scene of an empty hall way is deemed significant, see U.S. Patent Application 10/610,467, "Method for Detecting Short Term Unusual Events in Videos," filed by Divakaran et al. on June 30, 2003, incorporated herein by reference. Similarly, for a traffic video, a change from smooth motion to no motion can be indicative of an accident.

System Architectures

[037] Figure 6A shows a block diagram of a video encoding system for multiple videos in accordance with our invention. The videos 510 from multiple camera 601 are first NTSC decoded 602. The digitized videos are then input to the joint analysis circuit 600 to determine a temporal sampling rate 604 for each video.

[038] The temporal sampling rate 604 is determined according to a method that considers the presence of significant events in the video, compression efficiency and resource constraints. The method and constraints are described in greater detail below.

[039] The variable and non-uniform temporal sampling rate for each video is passed to a controller 610, which triggers the sampling circuit 604. This circuit is capable of non-uniformly sampling frames of each video input at time *t*. The sampling times $T_1(t)$, $T_2(t)$, $T_3(t)$ and $T_4(t)$ do not necessarily operate at the same rate as one another. These sampling times determine the combined frame rate across all videos.

[040] The frames of the sampled videos 520 are then stored in a frame buffer 620. The controller 610 selects 630 the input to be encoded 640, and the frames are read from the respective frame buffers and passed to a video encoder 640. From the controller, the video encoder can receive the following information 611: coding parameters to be used for encoding the video that were derived by the joint analysis 600, and identification information that can also be encoded in the bitstream itself. The identification can be encoded directly in the bitstream using an

MPEG-4 video object plane coding syntax described in ISO/IEC 14496-2, "Coding of audio-visual objects – Part 2: Visual," 2nd Edition, 2001.

[041] The system produces a unique bitstream 650 for each video input, which is then stored to a persistent memory 660 for later playback.

[042] With this system, a minimal number of reference frames are stored for each video, typically one or two frames. The reason for this low memory requirement is due to the nature of the video encoding process. In contrast to the prior art system by Ono, et al. shown in Figure 4, which encodes *GOP's* of data, the system according to this invention encodes *frames* of data. This is enabled through the video encoder 640 that can mark the different video inputs and produce independent output bitstreams 650. The controller 610 can select 630 the correct reference frames to be used for motion compensation of the current video being encoded. Neither of these features are part of the system described by Ono, et al. These features can provide a scalable system for encoding concurrently multiple videos.

[043] The system in Figure 6B is a slightly different configuration of the system in Figure 6A. The main difference is a single shared frame buffer 621.

[044] Figure 7 shows an extension of the core architectures of Figure 6A and 6B. The bitstreams stored in the persistent memory 660 can be transferred to an archival memory 703 for long-term storage. In this case, a transcoder 701 can be employed to further modify the bit rate, spatial resolution and/or temporal resolution of the stored bitstream. An analysis circuit 702 is used to configure and set the transcoding parameters. Any prior art transcoding techniques may be used,

such as those described by Vetro, et al. in "An overview of video transcoding architectures and techniques," IEEE Signal Processing Magazine, March 2003.

[045] Although the above architectures are shown for four input videos and one video encoder, this design allows the system to easily be scaled up to a greater number of input videos and video encoders, e.g., 16 inputs and 2 video encoders, or 64 inputs and 4 video encoders.

[046] Joint Analysis

[047] The factors affecting the operation of the joint analysis 603 include: a combined distortion incurred by skipping frames, the number of encoders integrated as part of the system, the number of input videos, the detection and classification of significant events in a scene, users preference to a particular camera input, a minimum temporal coding rate for each video, the allowable delay and required storage for analysis.

[048] The combined distortion that is incurred by skipped frames guides the operation of the joint analysis. In U.S. Patent Application Sn., 09/ 835,650 "Estimating Total Average Distortion in a Video with Variable Frameskip," filed by Vetro et al. on April 16, 2001. The distortion of a skipped frame is determined from two parts: a coding error due to quantization of the last reference frame, and a frame interpolation error due to the change in the video signal between the two time instants. That method only considers distortion in a single video. Here, the total combined distortion of all output video is minimized.

[049] In the following, we refer to the distortion caused by the coding error as the spatial distortion, and the distortion caused by the interpolation error as the temporal distortion. Because video recording systems typically encode all frames at the same quality, the problem of estimating distortion can exclude the spatial distortion of frames and focus only on the temporal distortion. It is emphasized that this is quite different from any formulation that has been considered in the prior art.

[050] The number of encoders that are integrated as part of the system determines the combined frame rate, i.e., the maximum number of frames that can be encoded per unit time. For instance, if a single encoder can encode 30 frames per second and there are four encoders integrated as part of the system, then the combined frame rate is 120 frames/second. For example, the combined frame rate can be allocated over the video as follows. Sixty surveillance video of static scenes can be sampled at 1 one frame per second each, while two video with activity can be sampled at 30 frames per second, for a total combined frame rate of 120 fps.

[051] For the purpose of formulating a problem based on this constraint, a value $N_{cap}(T)$ denotes the total number of frames per unit of time T. Depending on the number of input videos to the system and the frame rate of those videos, the average temporal sampling rate across all videos can be determined.

[052] The objective function for the joint analysis can be stated as follows:

$$\min \sum_i D_i \quad \text{such that } N_{coded}(T) \leq N_{cap}(T) \qquad (1)$$

where D_i is the temporal distortion of video i, and $N_{coded}(T)$ is the total number of frames to be coded per unit of time T. To maximize the utilization of the encoding capability, $N_{coded}(T) = N_{cap}(T)$.

[053] The temporal distortion from (1) in a time interval $[t, t+\tau]$ is expressed as follows:

$$D_i[t, t+\tau] = w_i \sum D_{skipped\_frames}[t, t+\tau] \qquad (2)$$

[054] The above equations sum the distortion of skipped frames in the given time interval and considers a multiplicative weighting factor for each video, w_i. The purpose of this weighting, and factors that influence the weight values are described below.

[055] As stated earlier, this formulation is novel in that only the temporal distortion is accounted for, but it is also novel in that there are no rate constraints for individual videos, as is very typical in formulations in the prior art. Rather, there is a constraint on the combined frame rate for all videos. This is the second novel aspect of this formulation, and to our knowledge, this feature does not appear in any prior art systems.

[056] The detection of a significant event in the scene could imply a binary operation to record whether the event has been detected, and cease recording if there is no significant event. This is quite a simple operation. However, unless the detector is very accurate and robust to noise, this is not a very good strategy because there is a chance that some significant events could be missed, and that some insignificant events will be recorded.

[057] Rather, it is preferred to include a weighting factor as part of the above objective function to be optimized. Given that the objective function is based on minimizing the combined distortion, we employ a weight of unity when there is no

significant event, and a larger weight, e.g., $w_i = \mathrm{w}_{\max}$, when there is an event. The value of w_{max} is adjustable. Large values encourage the joint analysis to code more frames in the video in which an event has been detected in order to minimize the combined distortion. When there is no event detected and $w_i = 1$, the default mode is for the objective function to rely purely on the distortion of a video to determine the coded frames. The weight, w_i, can take on values in the range $[1, \mathrm{w}_{\max}]$ given that the process used for event detection has the ability to distinguish the degree of significance for different events.

[058] An additional use of the weight, w_i, in the formulation of (1) is to express the preference of a particular video. For instance, if a video is of a critical area, e.g., an entry way, then w_i may be set larger than one. Similar to the detection of an event, this gives bias to always encode more frames from this input video. Note, the preference can also operate on the weighting that results from the event detection. In other words, the preference can add a constant value to the weight that is output from the event detection process.

[059] Rather than adding preference for a video, the weight can also be used to reduce the emphasis on a relative number of frames to be coded for a particular video. This can be achieved with weights less than one. The reason is that the distortion incurred by skipping frames from that input video accumulates slower, and fewer frames are coded.

[060] However, there can be a minimum temporal coding rate that is set by the system. This setting forces a frame to be coded for a particular input after some predetermined period has elapsed since the last frame was encoded. Setting the weight to zero for a particular input always force this minimum temporal rate.

[061] The final factor influencing the operation of the joint analysis 603 is the allowable delay (τ). If a larger delay can be tolerated, then the problem is solved considering a larger interval of time. Although lower distortions are possible with a larger time interval, there are two drawbacks to having this larger time interval. The first drawback is memory size. A larger window of time implies that more storage is needed to analyze incoming frames. The second drawback is computation. For a larger window of time, there are more possible solutions that need to be evaluated. This is a typical engineering trade-off that needs to be made during the system design. However, we note the flexibility that the system described by this invention offers.

[062] Breaking the problem down into a fixed time interval, *T*, and letting *N* denote the combined total number of possible frames that are encoded in this time interval for all input videos, the optimal solution to the above stated problem requires the system to evaluate $\binom{N}{N_{cap}}$ possible solutions.

[063] The solution with the minimum combined distortion over all videos for the given time interval is selected. It should be noted that each of the possible solutions is a combination of distortion calculations between two frames. If *M* is total number of input videos and *K* is the total number of frames within the period *T* for each video, then the total number of distinct distortions between pairs of frame that need to be calculated is $M \times \binom{K}{2}$.

[064] In an embodiment shown in Figure 8, the parallel input videos 901 to the joint analysis 800 are independently compressed frames, i.e., intra-frame

compressed videos. Examples of videos that are part of the intra-frame compressed category include videos whose frames are compressed using JPEG, JPEG 2000, MPEG-1/2 I-frames, MPEG-4 I-VOP's, (video-object-planes) or any other intra-frame compression technique.

[065] In a networked surveillance system, the camera inputs are compressed in this way to reduce bandwidth. At the same time, editing and review operations can still access each frame independently. Using intra-frame compression in this way is also quite common in broadcast studios, where a video is usually transmitted internally to different editing stations and the video undergoes a number of changes before the video is broadcast.

[066] In this embodiment, the joint analysis 800 still operates according to the objective function and constraints given in equation (1). However, in this case, the temporal distortion is determined directly from the compressed domain information in the intra-frames. This compressed domain information can include discrete cosine transform (DCT) coeeficients. Because frame differences and correlation among frames can be achieved quite efficiently from the compressed-domain, it is not necessary to fully decode the frames and perform the analysis from the reconstructed frames. Instead, the analysis 800 is performed on partially decoded frames.

[067] The output of the controller 610 operates as before, in which the sampling circuit 604 is triggered to non-uniformly sample the compressed frames of each video input at time *t*. As an advantage, the sampling rates $T_1(t)$, $T_2(t)$, $T_3(t)$ and $T_4(t)$ can be different. After the sampling rates have been determined, the intra-

compressed videos are sampled accordingly and the samples frames are recorded directly to the persistent memory 660.

[068] In an embodiment shown in Figure 9, input videos 901 to the joint analysis 900 are interdependently compressed frames, i.e., inter-frame compressed videos. Videos that fall under this category include videos whose frames are compressed using MPEG-1/2 P/B-frames, MPEG-4 P/B-VOP's, or any other technique that predicts information in the current frame from another past or future reference frame.

[069] In a networked camera system, the videos are compressed to a greater extent than possible with intra-frame compression. This is particularly true for surveillance systems, where the amount of temporal redundancy in mostly static scenes is very high. One main advantage of this embodiment is that a large number of cameras can share the same network and bandwidth. In this case, some of the frames serve as reference frames to predict other frames. Care must be taken to process these frames correctly.

[070] In this embodiment, the joint analysis 900 estimates the temporal distortion from the compressed video. One means to achieve this is based on techniques described in U.S. Patent Application Sn., 09/ 835,650 “Estimating Total Average Distortion in a Video with Variable Frameskip,” filed by Vetro et al. on April 16, 2001.

[071] The temporal distortion between frames i and k, $E\{\Delta^2 z_{i,k}\}$, is estimated by,

$$E\{\Delta^2 z_{i,k}\}=\sigma_{x_i}^2\sigma_{\Delta x_{i,k}}^2+\sigma_{y_i}^2\sigma_{\Delta y_{i,k}}^2, \qquad (3)$$

where $(\sigma_{x_i}^2, \sigma_{y_i}^2)$ represent the variances for x and y spatial gradients in frame i, and $(\sigma_{\Delta x_{i,k}}^2, \sigma_{\Delta y_{i,k}}^2)$ represent the variances for motion vectors between the two frames in the x and y direction.

[072] There, Vetro et al. note a practical difficulty when that technique is applied to *encoding* a video, and the goal is to estimate the distortion of different frameskip factors. The object is to select the factor that minimizes distortion. However, motion vectors for many possible frames are not yet available at the time of encoding, and it is computationally difficult to estimate the motion vectors.

[073] However, in the embodiment of Figure 9, motion vectors are directly available from the compressed videos 901. Spatial gradients, e.g., texture and edge information, in the reference frame can be determined directly from DCT coefficients, see Wang, et al., "Survey of compressed-domain features used in audio-visual indexing and analysis," Journal of Visual Communication and Image Representation, Volume 14, Issue 2, pp. 150-183, June 2003.

[074] The estimated distortion given by equation (3) as well as any other means to measure the temporal distortion can be used in accordance with this invention. It is an object of the invention to *minimize the combined distortion in all videos* **and** *satisfy a combined frame rate constraint.*

[075] Based on the outcome of the joint analysis 900, information on the frames to be dropped from the input videos are passed to the controller 910. The controller signals this information to temporal rate transcoders 920, one for each video.

[076] It is noted that the transcoding operation is significantly less complex than an encoding operation. During encoding, motion vectors must be determined. This is not the case for transcoding, when motion vectors are available from the input streams. Thus, the system can concurrently process many videos.

[077] Various temporal rate reduction transcoding techniques are described by Vetro et al., “An Overview of Transcoding Architectures and Techniques,” by Vetro, et al., IEEE Signal Processing Magazine, vol. 20, no. 2, pp. 18-29, March 2003.

[078] After the temporal rate for each video has been reduced accordingly, the compressed output streams 902 are directly recorded to the persistent memory 660.

[079] It should be understood that the input videos can concurrently include both compressed and uncompressed videos.

[080] Although the invention has been described by way of examples of preferred embodiments, it is to be understood that various other adaptations and modifications can be made within the spirit and scope of the invention. Therefore, it is the object of the appended claims to cover all such variations and modifications as come within the true spirit and scope of the invention.